Information, learning and falsification

David Balduzzi
Max Planck Institute for Intelligent Systems, Tübingen, Germany.
david.balduzzi@tuebingen.mpg.de

There are (at least) three approaches to quantifying information. The first, algorithmic information or Kolmogorov complexity, takes events as strings and, given a universal Turing machine, quantifies the information content of a string as the length of the shortest program producing it [1]. The second, Shannon information, takes events as belonging to ensembles and quantifies the information resulting from observing the given event in terms of the number of alternate events that have been ruled out [2]. The third, statistical learning theory, has introduced measures of capacity that control (in part) the expected risk of classifiers [3]. These capacities quantify the expectations regarding future data that learning algorithms embed into classifiers.

Solomonoff and Hutter have applied algorithmic information to prove remarkable results on universal induction. Shannon information provides the mathematical foundation for communication and coding theory. However, both approaches have shortcomings. Algorithmic information is not computable, severely limiting its practical usefulness. Shannon information refers to ensembles rather than actual events: it makes no sense to compute the Shannon information of a single string – or rather, there are many answers to this question depending on how a related ensemble is constructed. Although there are asymptotic results linking algorithmic and Shannon information, it is unsatisfying that there is such a large gap – a difference in kind – between the two measures.

This note describes a new method of quantifying information, effective information, that links algorithmic information to Shannon information, and also links both to capacities arising in statistical learning theory [4, 5]. After introducing the measure, we show that it provides a non-universal analog of Kolmogorov complexity. We then apply it to derive basic capacities in statistical learning theory: empirical VC-entropy and empirical Rademacher complexity. A nice byproduct of our approach is an interpretation of the explanatory power of a learning algorithm in terms of the number of hypotheses it falsifies [6], counted in two different ways for the two capacities. We also discuss how effective information relates to information gain, Shannon and mutual information.

Effective information

Any physical system, at any spatiotemporal scale, is an input/output device. For simplicity, we only model memoryless systems with finite input $\mathcal{X}$ and output $\mathcal{Y}$ alphabets. The probability that system $\mathfrak{m}$ outputs $y \in \mathcal{Y}$ given input $x \in \mathcal{X}$ is encoded in Markov matrix $p_{\mathfrak{m}}(y|x)$.

The effective information generated when system $\mathfrak{m}$ outputs y is computed as follows. First, let the *potential repertoire* $p_{unif}(X)$ be the input set equipped with the uniform distribution. Next, compute the *actual repertoire* via Bayes' rule

$$\hat{p}(X|y) := p\big(y|do(x)\big) \cdot \frac{p_{unif}(x)}{p_{\mathfrak{m}}(y)}, \tag{1}$$

where $p_{\mathfrak{m}}(y) = \sum_x p_{\mathfrak{m}}\big(y|do(x)\big) \cdot p_{unif}(x)$ and $do(\cdot)$ refers to Pearl's interventional calculus [7]. *Effective information* is the Kullback-Leibler divergence between the two repertoires

$$ei(\mathfrak{m}, y) := D\big[\hat{p}_{\mathfrak{m}}(X|y) \, \big\| \, p_{unif}(X)\big]. \tag{2}$$

For a deterministic function $f : \mathcal{X} \to \mathcal{Y}$, the actual repertoire and effective information are

$$\hat{p}_f(x|y) = \begin{cases} \frac{1}{|f^{-1}(y)|} & \text{if } f(x) = y \\ 0 & \text{else} \end{cases} \quad \text{and} \quad ei(f, y) = \log_2 |\mathcal{X}| - \log_2 |f^{-1}(y)|. \tag{3}$$

The support of the actual repertoire is the pre-image $f^{-1}(y)$. Elements in the pre-image all have the same probability since they cannot be distinguished by the function f. Effective information quantifies the size of the pre-image relative to the input set – the smaller ("sharper") the pre-image, the higher ei.

Algorithmic information

We show that effective information is a non-universal analog of Kolmogorov complexity. Given universal Turing machine T, the (unnormalized) *Solomonoff prior* probability of string s is

$$p_T(s) := \sum_{\{i \mid T(i)=s\bullet\}} 2^{-\text{len}(i)}, \tag{4}$$

where the sum is over strings i that cause T to output s as a prefix, where no proper prefix of i outputs s, and len(i) is the length of i. *Kolmogorov complexity* is $K(s) := -\log_2 p_T(s)$. Kolmogorov complexity is usually defined as the shortest program on a universal prefix machine that produces s. The two definitions coincide up to additive constant by Levin's Coding Theorem [1].

Replace universal Turing machine T with deterministic system $f : \mathcal{X} \to \mathcal{Y}$. All inputs have len$(x) = \log_2 |\mathcal{X}|$ in the optimal code for the uniform distribution on $\mathcal{X}$. Define the *effective probability* of y as

$$p_f(y) = \sum_{\{x \mid f(x)=y\}} 2^{-\text{len}(x)} = \begin{cases} \frac{|f^{-1}(y)|}{|\mathcal{X}|} & \text{if } y \in f(\mathcal{X}) \\ 0 & \text{else.} \end{cases} \tag{5}$$

Note that $p_f(y)$ is a special case of $p_{\mathfrak{m}}(y)$, as defined after Eq. (1). The effective distribution is thus a non-universal analog of the Solomonoff prior, since it is computed by replacing universal Turing machine T in Eq. (4) with deterministic physical system $f : \mathcal{X} \to \mathcal{Y}$.

In the deterministic case, effective information turns out to be $ei(f, y) = -\log_2 p_f(y)$, analogously to Kolmogorov complexity. Effective information is non-universal – but computable – since it depends on the choice of f.

Statistical learning theory

This section uses a particular deterministic function, learning algorithm $\mathfrak{L}_{\mathcal{F},\mathcal{D}}$, to connect effective information and the effective distribution to statistical learning theory.

Given finite set $\mathcal{X}$, let *hypothesis space* $\Sigma_{\mathcal{X}} = \big\{\sigma : \mathcal{X} \to \pm 1\big\}$ contain all labelings of elements of $\mathcal{X}$. Now, given a set of functions $\mathcal{F} \subset \Sigma_{\mathcal{X}}$ and unlabeled data $\mathcal{D} \in \mathcal{X}^l$, define *learning algorithm* (empirical risk minimizer)

$$\mathfrak{L}_{\mathcal{F},\mathcal{D}} : \Sigma_{\mathcal{X}} \longrightarrow \mathbb{R} : \hat{\sigma} \mapsto \epsilon = \min_{f \in \mathcal{F}} \frac{1}{l} \sum_{k=1}^{l} \mathbb{I}\left[f(d_k) \neq \hat{\sigma}(d_k)\right]. \tag{6}$$

The learning algorithm takes a labeling of the data as input and outputs the empirical risk of the function that best fits the data. We drop subscripts from the notation $\mathfrak{L}$ below.

Define empirical VC-entropy in [3]) as $\mathcal{V}(\mathcal{F}, \mathcal{D}) := \log_2 |q_{\mathcal{D}}(\mathcal{F})|$ where $q_{\mathcal{D}} : \mathcal{F} \to \mathbb{R}^l : f \mapsto \big(f(d_1) \ldots f(d_l)\big)$. Also define empirical Rademacher complexity as

$$\mathcal{R}(\mathcal{F}, \mathcal{D}) = \frac{1}{|\Sigma|} \sum_{\sigma \in \Sigma} \left[\sup_{f \in \mathcal{F}} \frac{1}{l} \sum_{k=1}^{l} \sigma(d_k) \cdot f(d_k)\right].$$

These capacities can be used to bound the expected risk of classifiers, see [8, 9] for details. The following propositions are proved in [5]:

Proposition 1 (effective information "is" empirical VC-entropy)**.**

$$ei(\mathfrak{L}, 0) = -\log_2 p_{\mathfrak{L}}(0) = l - \mathcal{V}(\mathcal{F}, \mathcal{D})$$

Proposition 2 (expectation over $p_{\mathfrak{L}}(\epsilon)$ "is" empirical Rademacher complexity)**.**

$$\mathbb{E}[\epsilon \mid p_{\mathfrak{L}}] = \sum_{\epsilon \in \mathbb{R}} \epsilon \cdot p_{\mathfrak{L}}(\epsilon) = \frac{1}{2}\big(1 - \mathcal{R}(\mathcal{F}, \mathcal{D})\big)$$

Thus, replacing the universal Turing machine with learning algorithm $\mathfrak{L}_{\mathcal{F},\mathcal{D}}$ we obtain that our analog of Kolmogorov complexity, the effective information of output $\epsilon = 0$, is essentially empirical VC-entropy. Moreover, the expectation of the analog of the Solomonoff distribution is essentially Rademacher complexity.

The two quantities $ei(\mathfrak{L}, 0)$ and $\mathbb{E}[\epsilon \,|\, p_{\mathfrak{L}}]$ are measures of explanatory power: as they increase, so expected future performance improves. By Eq. (3), the effective information generated by $\mathfrak{L}$ is

$$ei(\mathfrak{L},0) = \underbrace{\log_2|\Sigma|}_{\text{total \# hypotheses}} - \underbrace{\log_2|\mathfrak{L}^{-1}(0)|}_{\text{\# hypotheses } \mathfrak{L} \text{ fits}} = \Big(\text{\# hypotheses } \mathfrak{L} \text{ falsifies}\Big), \tag{7}$$

where hypotheses are counted after logarithming. Effective information, which relate to VC-entropy, counts the number of hypotheses the learning algorithm falsifies when it fits labels perfectly, without taking into account how often they are wrong. Similarly, see [5] for details, the expectation is

$$\textstyle\sum_\epsilon p_{\mathfrak{L}}(\epsilon)\cdot\epsilon = \sum_\epsilon \Big(\text{fraction of hypotheses } \mathfrak{L} \text{ falsifies}\Big)\cdot\Big(\text{on fraction } \epsilon \text{ of the data}\Big). \tag{8}$$

Expected ϵ, which relates to Rademacher complexity, looks at the average behavior of the learning algorithm, averaging over the fractions of hypotheses falsified, weighted by how much of the data they are falsified on.

The bounds proved in [3, 8, 9], which control the expected future performance of the classifier minimizing empirical risk, can therefore be rephrased in terms of the number of hypotheses falsified by the learning algorithm, Eqs (7) and (8), suggesting a possible route towards rigorously grounding the role of falsification in science [6].

Shannon information

We relate effective information to Shannon and mutual information.

Suppose we have model $\mathfrak{m}$ that generates data $d \in D$ with probability $p_{\mathfrak{m}}(d|h)$ given hypothesis $h \in H$. For prior distribution $p(H)$ on hypotheses, the information gained by observing d is

$$D\big[p_{\mathfrak{m}}(H|d) \,\big\|\, p(H)\big]. \tag{9}$$

Kullback-Leibler divergence $D[p\|q]$ can be interpreted as the number of Y/N questions required to get from q to p. Thus, Eq. (9) quantifies how many Y/N questions the model answers about the hypotheses using the data.

Effective information, Eq. (2), quantifies the information gained when physical system $\mathfrak{m}$ outputs y. Rather than inferring on hypotheses, the system, by producing an output, specifies probabilistic constraints on what its input must have been. Effective information uses the uniform (maximum entropy) prior since any other prior would insert additional data not belonging to the system – the prior is something else, *on top of* $\mathfrak{m}$. However, this restriction is not essential and will be dropped for the remainder of this section.

Consider the following scenario. We have X and X' are isomorphic, and a deterministic physical system $c : X \to X'$ that *copies* its inputs, mapping $x_k \mapsto x'_k$ for example. Given prior $p(X)$, the effective information generated is

$$ei\big(p(X), c, x'_k\big) := D\big[p_c(X|x'_k) \,\big\|\, p(X)\big] = D\big[\delta_{x_k} \,\big\|\, p(X)\big] = -\log_2 p(x_k),$$

the surprise of x_k. It follows that Shannon information is expected effective information

$$H(X) = \mathbb{E}\Big[ei\big(p(X), c, x'_k\big) \,\Big|\, p_c(X')\Big].$$

More generally, if we are given noisy memoryless channel $\mathfrak{m}$ from $\mathcal{X}$ to $\mathcal{Y}$ with distribution $p(X)$ on X, then mutual information is the expectation

$$I(X;Y) = \mathbb{E}\Big[ei(p(X), \mathfrak{m}, y) \,\Big|\, p_{\mathfrak{m}}(Y)\Big],$$

where $p_{\mathfrak{m}}(y) = \sum_x p_{\mathfrak{m}}\big(y|do(x)\big) \cdot p(x)$ is the effective distribution on Y. Thus, Shannon and mutual information are simply averages of effective information, our non-universal analog of Kolmogorov complexity.

Finally, interpreting effective information as information gain, Eq. (9), and combining with results from the previous section shows that $(l -$ empirical VC-entropy$)$ is the information we gain about the set Σ_X of hypotheses when told that learning algorithm $\mathfrak{L}_{\mathcal{F},\mathcal{D}}$ fit the labeled data perfectly.

Discussion

This note starts from the observation that all physical systems classify inputs and thereby generate information. A deterministic physical system $f : X \to Y$ implicitly categorizes its inputs by assigning them to outputs: the category assigned to output y is the set of inputs in the pre-image $f^{-1}(y) \subset X$. The intuition carries through in the probabilistic case after replacing pre-images with actual repertoires. Effective information then quantifies the sharpness of categories: the sharper a category, the more informative the corresponding output. Alternatively, effective information quantifies causal dependencies: outputs with high ei are extremely sensitive to changes in the input.

Effective information is a concrete, computable analog of Kolmogorov complexity. The Kolmogorov complexity of a string quantifies the "work" required to produce it; roughly, the length of the programs that output it. Since universal Turing machines require infinite storage space and are therefore impossible to construct, it is unclear how relevant they are to processes actually occurring in nature. Effective information substitutes a deterministic model of a physical system in place of the universal Turing machine, and quantifies the "work" required to produce an output as the number of Y/N decisions required to choose it.

Both Shannon and mutual information arise as expectations of effective information after tweaking to get rid of the uniform prior. The difference between Kolmogorov complexity and Shannon information reduces to: (i) replacing a universal Turing machine with a specific system (channel) and (ii) computing the average information gain over all outputs, rather than a single one.

When the physical process under consideration is empirical risk minimization, the effective information it generates contributes to bounds on expected risk. In particular, the work (the number of Y/N decisions) required to fit data $\mathcal{D}$ using functions in $\mathcal{F}$ essentially is the empirical VC-entropy. Since finding the optimal classifier in $\mathcal{F}$ requires computing $\mathfrak{L}_{\mathcal{F},\mathcal{D}}$ in some way or another, thereby implementing it physically, it follows that the effective information generated while fitting data has implications for the future performance of classifiers, see [3, 8, 9].

Effective information and the expected risk over the effective distribution also provide new interpretations of VC-entropy and Rademacher complexity in terms of falsifying hypotheses, see Eq. (7) and (8) – and also [10] for a comparison of falsification with VC-dimension. Viewing empirical risk minimization as a physical process that classifies hypotheses according to fit ϵ thus directly links VC-entropy and Rademacher complexity with Popper's proposal that the power of a scientific theory lies in how many hypotheses *it rules out*, rather than the amount of data it explains [6].

The links with Kolmogorov complexity, learning theory, information gain and falsification shown above suggest it is worth investigating whether the effective information generated while optimizing quantities other than empirical risk (e.g. margins) has implications for future performance.

Acknowledgements. I thank Samory Kpotufe and Pedro Ortega for useful discussions.

References

[1] Li M, Vitányi P (2008) An Introduction to Kolmogorov Complexity and Its Applications. Springer.

[2] Shannon C (1948) A mathematical theory of communication. Bell Systems Tech J 27:379–423.

[3] Vapnik V (1998) Statistical Learning Theory. John Wiley & Sons.

[4] Balduzzi D, Tononi G (2008) Integrated Information in Discrete Dynamical Systems: Motivation and Theoretical Framework. PLoS Comput Biol 4:e1000091. doi:10.1371/journal.pcbi.1000091.

[5] Balduzzi D (in press) Falsification and Future Performance. In: Proceedings of Solomonoff 85th Memorial Conference. Springer Lecture Notes in Artificial Intelligence.

[6] Popper K (1959) The Logic of Scientific Discovery. Hutchinson.

[7] Pearl J (2000) Causality: models, reasoning and inference. Cambridge University Press.

[8] Boucheron S, Lugosi G, Massart P (2000) A Sharp Concentration Inequality with Applications. Random Structures and Algorithms 16:277–292.

[9] Bousquet O, Boucheron S, Lugosi G (2004) Introduction to Statistical Learning Theory. In: Bousquet O, von Luxburg U, Rätsch G, editors, Advanced Lectures on Machine Learning, Springer. pp. 169–207.

[10] Corfield D, Schölkopf B, Vapnik V (2009) Falsification and Statistical Learning Theory: Comparing the Popper and Vapnik-Chervonenkis Dimensions. Journal for General Philosophy of Science 40:51–58.